

Reply to “Comment on ‘A Scaling law beyond Zipf’s law and its relation to Heaps’ law’ ”

Francesc Font-Clos^{1,2} and Álvaro Corral^{1,2}

¹*Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain.*

²*Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain*

In Ref. [1], Bernhardsson, da Rocha, and Minnhagen proposed that the distribution of word frequencies in a text or collection of texts (of the same author) changes with text length as

$$D_L(k) = A \frac{e^{-k/(c_0 L)}}{k^{\gamma(L)}}, \quad (1)$$

where k is the absolute frequency (number of tokens) of the different words (word types), L is text length in number of tokens (M in Ref. [1]’s notation), $D_L(k)$ is the probability mass function of k (i.e., the distribution of word frequencies), $\gamma(L)$ is the power-law exponent, c_0 is a scale parameter (independent on L), and A is a normalizing constant. The key ingredient of Bernhardsson et al.’s approach to model the change of $D_L(k)$ with L is the explicit dependence of the exponent γ on text length L , decreasing with increasing L .

Alternatively, in Ref. [2], we argue that the variability of the statistics of words in a text with its size is better explained by a scaling law,

$$D_L(k) = \frac{1}{LV_L} g(k/L), \quad (2)$$

where V_L is the size of vocabulary (number of different words, i.e., word types), and $g(x)$ is a undefined scaling function, independent of text size.

Now, Yan and Minnhagen [3] claim that our scaling law is “fundamentally impossible” and “fundamentally incorrect”. We summarize the points of these authors to make it clear that their criticism is essentially irrelevant. First, in Fig. 1, they find that our scaling law does not hold for $k = 1$. Second, in Fig. 2 they show that our scaling does not work well for, let us say, $k \leq 10$. Third, it is argued that a “Randomness view”, based in the concepts of “Random Group Formation”, “Random Book Transformation”, and “Metabook” predicts the right form of $D_L(k)$, which is that of Ref.[4].

It is obvious that the first and second criticisms of Yan and Minnhagen are irrelevant, as they simply imply that our scaling law can only be valid beyond the low-frequency limit, so,

$$D_L(k) = \frac{1}{LV_L} g(k/L), \text{ for } k > 10.$$

This is not surprising at all, as it is well known in statistical physics that scaling laws hold asymptotically. It is remarkable that, for texts, scaling is attained after the first decade in frequencies. It is also remarkable that, despite the fact that Yan and Minnhagen stretch the scaling hypothesis up to very small fragments of texts (212473/500 \simeq 400 tokens, for the case of *Moby-Dick*), the scaling law still is fulfilled reasonably well, beyond the first decade in k . Naturally, the appropriate way to further test the validity of our scaling law is in the opposite way, analyzing larger and larger texts.

To make our point more clear, in Fig. 1 we present the same data as in Fig. 2 of Ref. [3], but adding symbols for $k = 1 \dots 5$ (instead of only lines, as in Ref. [3]). It is apparent that even in the extreme case of $n = 500$, the scaling law only fails for very small frequencies. Additionally, in Fig. 2 we perform the data collapse associated to our scaling law for the case of *Harry Potter*, presented in Ref. [3] as a counter-example to our scaling law. As it is shown, the collapse is excellent: after proper rescaling, all curves collapse into a single, length-independent function, even for very small frequencies.

So, the empirical facts are clear: a scaling law gives a very good approximation for the distribution of word frequencies in the range $k > 10$. If the “Randomness view” hold by Yan and Minnhagen is valid, then it must contain in some limit the scaling law. If not, their theory is wrong. As a final remark, let us state that although curve fitting is a very honorable approach in science (when done correctly [5]), our scaling approach has nothing to do with that, contrary to Yan and Minnhagen’s claims.

In summary, the objections raised by Yan and Minnhagen are too weak to justify the publication of a comment to our work.

-
- [1] S. Bernhardsson, L. E. Correa da Rocha, and P. Minnhagen. The meta book and size-dependent properties of written language. *New J. Phys.*, 11:123015, 2009.
 - [2] F. Font-Clos, G. Boleda, and A. Corral. A scaling law beyond Zipf’s law and its relation to Heaps’ law. *New J. Phys.*, 15:093033, 2013.
 - [3] X.-Y. Yan and P. Minnhagen. Comment on ‘A scaling law beyond Zipf’s law and its relation to Heaps’ law’ [arxiv:1303.0705], 2014.
 - [4] S. K. Baek, S. Bernhardsson, and P. Minnhagen. Zipf’s law unzipped. *New J. Phys.*, 13(4):043004, 2011.
 - [5] A. Deluca and A. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.*, 61:1351–1394, 2013.

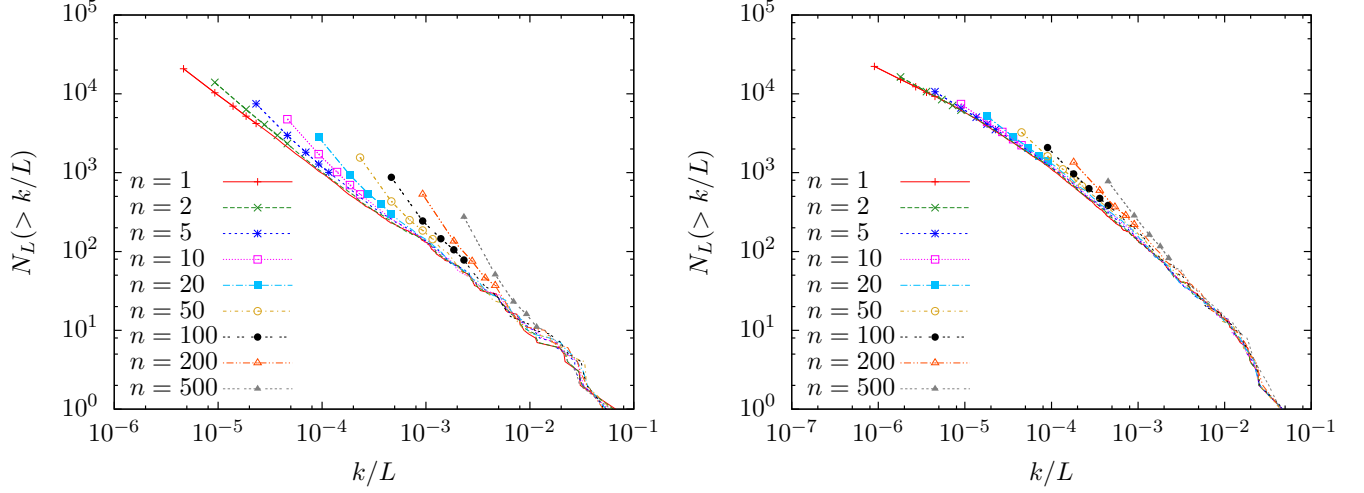


FIG. 1. The total number of words N_L with a relative frequency greater than or equal to k/L , for varying $L = L_{\text{tot}}/n$. We have taken the same books as in Ref. [3], *Moby-Dick* (left) and *Harry Potter* (right), exactly reproducing panels (a) and (b) of Fig. 2 in Ref. [3], but also including some additional values of n . Lines are drawn for all k , but symbols are drawn only for $k = 1 \dots 5$, showing that deviations from the scaling law are always in the regime of very low frequencies, as expected due to discreteness effects.

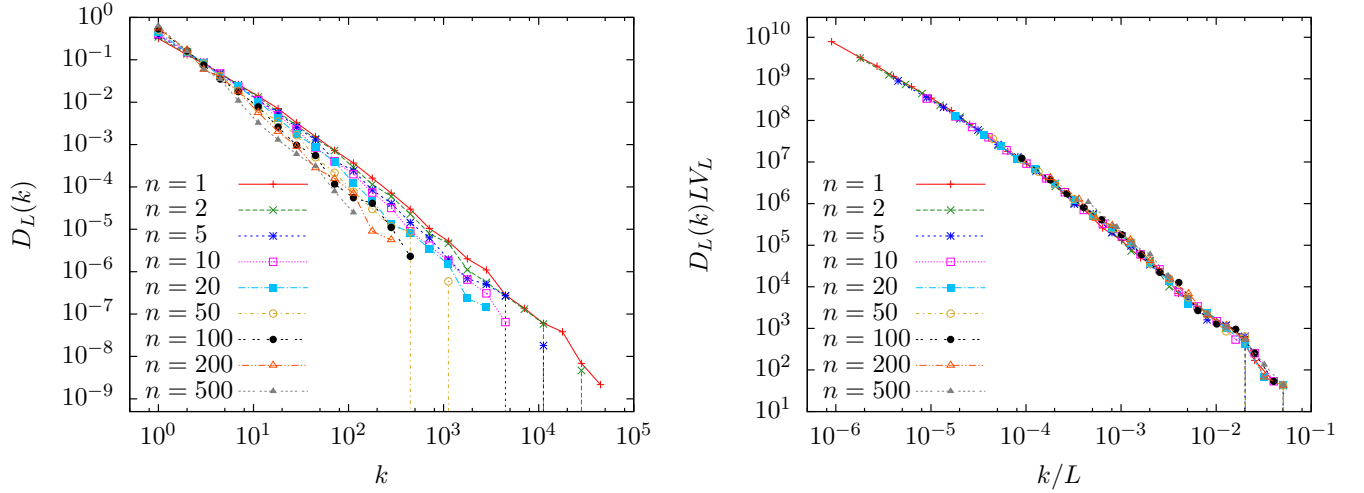


FIG. 2. **Left:** The probability mass function $D_L(k)$ of the absolute frequency k , for varying subsets of length $L = L_{\text{tot}}/n$ of *Harry Potter*, displaying a seeming change of shape. **Right:** Same, but plotting $D_L(k)LV_L$ versus k/L , as proposed in Ref. [2] and stated here in Eq. (2). All curves collapse into a single, length-independent scaling function $g(k/L)$, in agreement with Eq. (2), and at odds with Eq. (1): a length-dependent exponent in $D_L(k)$, as proposed by Yan and Minnhagen, is not compatible with the data collapse shown in the figure.